

午前3時の データマイニング

第2回：三河屋サブちゃんとマイニングする!?

奥田 真一郎
OKADA MASAICHIRO



Illustration by T. Chikashi

♪♪うーらのは一たけでボチがなくー。
正直じーさん堀ったらばー、おーばー
あんこーばあんが、ざーつくざーつく
ざっくざく♪♪ 皆さん、おはようご
ざいます。「午前3時のデータマイニン
グ」の時間です。先月に引き継ぎ、今月
も楽しくデータマイニングを学んでま
いりましょう。今回は、データマイニン
グの理解をさらに深めるため、データ
マイニングと既存の分析手法との違い
を見ながら、データマイニングの特徴
と実践の方向性を考えていくことにし
ましょう。

三河屋サブちゃんと クロスセリングを考える

前回は、データマイニングとは何か
についてその概要を見てきました。あ
れから1ヶ月経ちましたが、皆さん覚
えていますか？ マイニングの応用例と
して「クロスセリング」という言葉が
出てきましたね。何のこと？ と忘れてし
まった方のために、まずは簡単に復習
をしておきましょう。

クロスセリングとは、いわゆる「合わ
せ売り」のことです。ここで三河屋サブ
ちゃんの例を考えてみます。皆さん、三
河屋のサブちゃんってご存知ですか？
そうです。あの「サザエさん」に登場す
る、磯野家御用達の酒屋「三河屋」のお

兄さんです。いつも元気に磯野家の勝
手口から「ちわー！ 三河屋で一す！」
と注文のあった品物の配達にやってき
ます。時には、御用聞きにうかがうとい
う、まさに顧客密接型のマーケティング
戦略を展開しています。

例えば、サブちゃんが磯野家から注
文を受け、「みりん」を配達にきたとし
ます。すると、サブちゃんはすかさず、
サザエさんに対して「毎度ありがどう
ございます。他にご利用はございません
か？」との決まり文句に続けて、「い
やー。今日は新潟の蔵元からいいお酒
が入っていますよ！」などとさりげなく
購買を促します。これが、いわゆる
「クロスセリング」です。

ここで問題になるのは、どの商品
を選んでサザエさんに薦めるか？ とい
うことです。サザエさんは、大変「賢い」
(?消費者の1人ですので、単純に今日
売りたい商品をただ紹介しただけでは
「今日は結構よ！ おほほほほ」と軽くあ
しらわれてしまうでしょう。また、あま
りニーズのない商品をしつこく薦めて
も、逆にうっとおしがられるのが人情
です。

やはり、ここでは「そろそろ醤油が切
れていませんか？」とか「マスオさんに
ぴったりのワインが入りましたよ」な
ど、タイムリーかつお客様のニーズに

合った情報を提供して磯野家を攻めたい
ところです。しかし、サブちゃんはい
ささか先生のお宅など、他にもたくさ
んのお得意様を抱えています。すべての
ニーズを把握するのは大変ですね。

そこで、データマイニングが登場す
るわけです。その方法はいろいろ考え
られますが、まずは取引の履歴データ
を活用し、従来の配達作業からプラス
アルファのビジネスチャンスを開くこ
とを考えてみましょう。

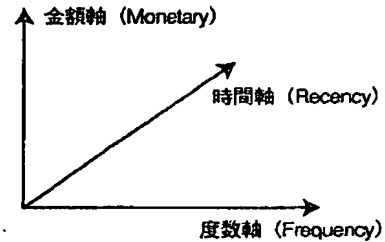
検索・集計を超えろ！

取引データとは、裏のような「いつ」
「誰が」「何を」「いくらで」「どれくらい」
買ったかというようなシンプルなデー
タです。ここでは、個人情報ではなく世
帯情報という視点から見ることにしま
しょう。

ただし、ここで気を付けなければなら
ないのは、前回でも触れたように
データをやみくもにこねくり回しても
仕方がないということです。データを
分析する前には、必ず「何を知りたいの
か」「どのようにすれば分かるのか」を
準備しなければなりません。さもない
ければ、たちまち「データジャングル」の
中に迷い込んでしまいます。では、何を
知ればサブちゃんはハッピーに(販売数
を伸ばせる)なれるのでしょうか？

Data Mining at 3:00 A.M.

●RFMを軸に顧客の総合指標化



顧客ID	日付	購入品目	数量	単価	金額
0000001	1999/10/1	ビール	12	350	4200
0000001	1999/10/2	みりん	1	400	400
0000001	1999/10/2	焼酎	1	300	300
0000002	1999/10/1	ビール	6	350	2100
0000002	1999/10/1	赤ワイン	1	1500	1500
0000003	1999/10/1	焼酎	1	300	300
0000003	1999/10/1	ビール	6	350	2100
0000003	1999/10/2	日本酒	1	1800	1800
0000003	1999/10/2	しょうゆ	1	400	400
0000004	1999/10/1	ビール	5	350	1750
0000004	1999/10/1	赤ワイン	1	1500	1500
0000004	1999/10/2	ビール	24	350	8400
0000004	1999/10/2	焼酎	1	300	300

●「焼酎」を購入を多く繰り返して行
るお客さんは誰と誰と最近に
買ったかを教えてください。

●10月15日現在のうちにどれだけ
買ってくれるお客さんは誰か。

●10月15日(金曜日)のうちにどれ
買ってくれるお客さんは誰か。

これらの質問は、次の3つの軸を
軸に考えます。

- ・時間軸(Recency)
- ・度数軸(Frequency)
- ・金額軸(Monetary)

この3つの軸を軸に、お客さんご
とごとの購買履歴を分析し、どの
お客さんがどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に

●10月15日現在のうちにどれだけ
買ってくれるお客さんは誰か。
●10月15日(金曜日)のうちにどれ
買ってくれるお客さんは誰か。

1. どの程度にどの程度にどの程度に

この3つの軸を軸に、お客さんご
とごとの購買履歴を分析し、どの
お客さんがどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に
どの程度にどの程度にどの程度に

●10月15日現在のうちにどれだけ
買ってくれるお客さんは誰か。
●10月15日(金曜日)のうちにどれ
買ってくれるお客さんは誰か。

では、10月15日(金曜日)のうちに
どれ買ってくれるお客さんは誰か。

1. どの程度にどの程度にどの程度に

●10月15日現在のうちにどれだけ
買ってくれるお客さんは誰か。
●10月15日(金曜日)のうちにどれ
買ってくれるお客さんは誰か。

では、10月15日(金曜日)のうちに
どれ買ってくれるお客さんは誰か。

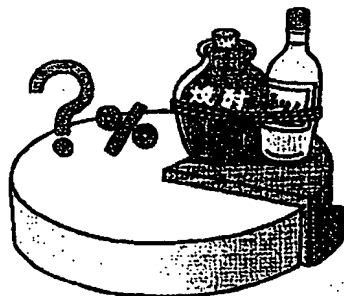


検索の処理だけでは、簡単に答えを出すことができません。しかし、この情報をつかめば、お客様にお薦めする商品が浮かび上がってきますね。ではどうすればよいのでしょうか？

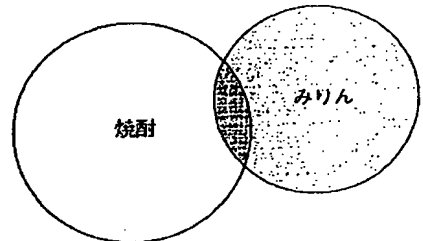
これぞマイニングの出番です。あらゆる組み合わせを探索し、件数の多い順にレポートする、ただそれだけの作業のような気がしますが、商品の組み合わせは2品目に限ったことではありません。取り扱い商品の中からすべての組み合わせを取り出すとなると、組み合わせの数は膨大になり、その中から有効な組み合わせを見つけるのは大変なことです。

データマイニングでは、この組み合わせの抽出を「アソシエーションルール」(Association Rule)と呼びます。アソシエーションは、図2のように、サポート(Support)とコンフィデンス(Confidence)という2つの基準により、特徴的な組み合わせを評価するものです。

サポートとは、全顧客(ユニークなIDの数)の中で、その商品の組み合わせで購入した人が何人(ここでは何世帯)いるのかを表わした指標です。例えば、「焼酎」と「みりん」の組み合わせで購入した世帯数は、全体の何パーセントになるのか、というようなものです。



数値例：
全顧客世帯数=100世帯
焼酎かつみりん購買世帯数=20世帯
焼酎購入数=50世帯
みりん購入世帯数=40世帯



焼酎購入世帯が同時に、みりんも購入する→ $20/50=40\%$ (コンフィデンス)
全顧客世帯の中で焼酎購入者が、みりんも購入する→ $20/100=20\%$ (サポート)

図2：アソシエーションはサポートとコンフィデンスにより抽出できる

これに対してコンフィデンスとは、ある商品を購入した顧客の中で、同時に別の商品と組み合わせで購入する割合を言います。例えば、「焼酎」を購入した世帯の中で、「みりん」も同時に購入した世帯の割合を指します。

これらの方法によって求められた組み合わせの中から、びっくりするようなおもしろいルールを探し出せることがデータマイニングの魅力の1つといえるでしょう。ここで紹介したアソシエーションルールの抽出は、データマイニングの最も特徴的な手法であり、非常に簡単なものですのでいろいろ応用できます。例えば、前回紹介したビールとおむつの関連も、このサポートとコンフィデンスにより浮かび上がってきたアソシエーションルール、「ビール→(ならば)おむつ」です。

OLAPとの違いは何か

いかがでしたか？ 網野家攻路の糸口と、マイニングの特徴が浮かび上がってきたでしょうか。では、ここでデータマイニングと比較すべきデータ分析のアプローチについてちょっと触れておきましょう。

皆さん、OLAPってご存知ですか？

意外にOLAP(オーラップと読みます)とデータマイニングを混同されている方が多いようなので、今一度整理しておきましょう。OLAPとは、On-Line Analytical Processing の略で、リレーショナルデータベース(RDBMS)の弱点を補完するために、米国のコード博士が提唱した12のルールに基づいて発展した分析手法です(図3)。

日本では、「多次元データ分析」と訳されていますが、データ解析の手法としての視点からいえば、実はOLAPはデータの「集計」にすぎません。多次元データ分析と聞くと、なんとなくすごそうな感じがしますが、実際には「オンライン集計処理」「集計済み結果表示」の仕組みとして分類できます。集計結果は、クロス集計のイメージやグラフで表現できます。

当然ですが、オンライン上でクロス集計結果を見ることができるとしても、1つ1つのデータを眺めるのは人間の役割です。サザエさんの例では、サブちゃん自身があらゆるクロス集計表を眺めながら考える姿をイメージすれば、雰囲気は分かると思います。

また、OLAPでは紙にプリントアウトしたクロス集計表とは異なり、ドリ

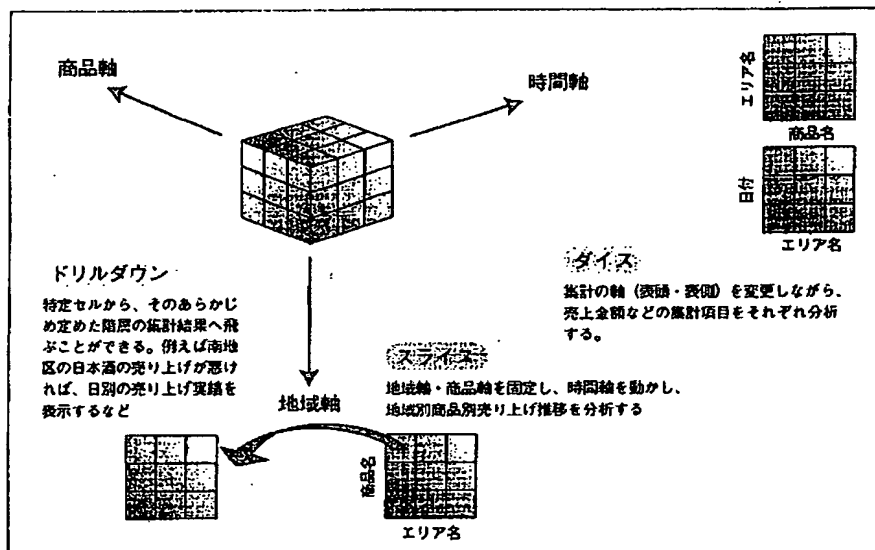


図3: OLAP の機能概要

ルダウン(Drill Down)やスライス&ダイス(Slice & Dice)という機能により、必要な集計項目を自由に変えたり、あらかじめ定めた集計の次元を切り替えながら結果を見ることができます。必要な項目を定めれば、常に結果をモニタリングできるので、大変便利な機能と言えるでしょう。

しかし、私たちはおもちゃのルービックキューブでさえ、組み変えているうちに元に戻せなくなってしまいます。データ探索にOLAPを用いるとい

うことは、実質的に集計・検索を繰り返すことと同じです。このため、データのジャングルに迷い込んでしまう可能性も高いと言えるでしょう。

これに対してマイニングでは、先のアソシエーションのように、注目すべき組み合わせをレポートしてくれます。探索そのものにかかる時間を短縮できるため、その結果を考察するという本来の業務に多くの時間を費やすことができるのです。

マイニングから得られた結果は、OLAPの次元として採用することも可能です。また、ドリルスルーといった詳細データをクロス表の特定セルから参照するような機能と組み合わせることもできます。マイニングとOLAPは、目的に合わせてそれぞれを使い分けていくことが大切です。

たかがアソシエーション、されどアソシエーション

今月はデータマイニングの特徴を三

河屋のサブちゃんと一緒に考えてきました。アソシエーションルール、クロスセリングについては、先月に引き続き細かく見てきましたので、お分かりいただけたと思います。

このほかにもクロスセリングは、商品の組み合わせだけでなく、顧客属性や天気など、さまざまな外部情報も組み込むことができます。例えば、「男性 & 30代 & ライトビール → 赤ワイン」など、より複雑な連関ルールを探索することが可能となるわけです。そして、このようなアソシエーションルールの発見によるクロスセリングは、小売業だけで応用されているものではありません。例えば、不良債権の原因を探索するなど、金融商品の分析もできますし、Webのアクセスログを利用すれば、サイトとサイト、ページとページの連関を発見することができます。また、医薬品の市販後の調査など、副作用をもたらす処方

の組み合わせの探索もできるでしょう。このように、アソシエーションにデータをどのように与えていくかのアイデア一つによって、いろいろなことが見えてきます。埋もれた大量のデータにはまだまだダイヤモンドが眠っているのです。ぜひ、皆さんもアソシエーションの身近な応用例を考えてみてください。きっと、サブちゃんも喜びますよ！ どうやら今月は、裏の畑で「たま」も鳴いているようですね。

奥山 真一郎 (おくやましんいちろう)
国際大学修了後、SAS インスティテュートジャパン入社。データサイエンスグループにて、統計解析、データマイニングのコンサルテーションに従事する。現在は「孤高のDataMiner」として、データとの付き合い方を模索中。
E-Mail: FZH04331@nifty.ne.jp

